The Decomposition of Similarity-Based Diversity and its Bias Correction

Eric Marcon, Zhiyi Zhang, Bruno Hérault

▶ **To cite this version:**

Eric Marcon, Zhiyi Zhang, Bruno Hérault. The Decomposition of Similarity-Based Diversity and its Bias Correction. 2014. hal-00989454v2

HAL Id: hal-00989454

https://agroparistech.hal.science/hal-00989454v2

Preprint submitted on 20 May 2014 (v2), last revised 18 Mar 2015 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Decomposition of Similarity-Based Diversity and its Bias Correction

Eric Marcon[1]*, Zhiyi Zhang[2], Bruno Hérault[3]

Abstract
Similarity-based diversity defined by Leinster and Cobbold (2012) encompasses classical measures of diversity, including the number of species, Shannon and Simpson diversity, and also phylogenetic and functional diversity. We derive two estimators to allow applying it to real, often under-sampled data, and its decomposition into alpha, beta and gamma diversity when an assemblage of communities is considered.

Keywords
Entropy, Biodiversity, Phylogenetic diversity, Functional diversity

[1] *AgroParisTech, UMR Écologie des Forêts de Guyane, BP 709, F-97310 Kourou, French Guiana.*
[2] *Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223*
[2] *CIRAD, UMR Écologie des Forêts de Guyane, F-97379 Kourou Cedex, France*
***Corresponding author**: Eric.Marcon@ecofog.gf

Contents

1 Introduction 1
2 Methods 2
2.1 Notations . . . 2
2.2 Definition of ${}^qD^Z$ and ${}^qH^Z$ . . . 2
2.3 Estimation bias correction . . . 3
Chao-Shen estimator • Alternative estimator
2.4 Decomposition of entropy . . . 4
2.5 Decomposition of diversity . . . 5
3 Results 5
3.1 Test of bias correction . . . 5
3.2 Partitioning functional diversity . . . 6
4 Discussion and Conclusion 6
References 7

1. Introduction

It is still surprising that a so widely used ecological concept, such as the concept of diversity, is still so debated. Ecology is rich of such recurring debates on concepts (*e.g.* the "ecological resilience" concept [1], or the "functional trait" concept [2]) that could nevertheless be seen as part of a fundamental theoretical corpus. This may be why one may consider that ecology is still a "young" science [3]. Practically, the notion of diversity is more or less consensual among field ecologists. Very roughly, the diversity of an ecological system is even higher than the individuals in the system are "different". Two ecological systems are even more different than their individuals are, too. Problems arise when we move from words to mathematical quantification. Entropy, viewed as the average surprise provided by the data [4,5], paved the way of a coherent theoretical framework [6] that respects a set of meaningful axioms [7,8] including discarding species identities, continuity relatively to the probabilities of occurrence of species and the Pigou-Dalton [9] property (replacing an individual of a more abundant species by an individual of a less abundant species increases diversity). Generalized entropy, namely HCDT entropy [10–12], allows moving the cursor, the order of diversity q, from rare to abundant species. Conversion of entropy into Hill numbers [13–15] provides the effective numbers, *i.e.* the number of equally-frequent entities (*e.g.* species or communities) which would yield the same diversity value as the data. The product of the effective number of communities [16], namely β diversity, by α diversity (the average diversity of communities) is equal to the γ diversity of a mixture of communities (*i.e.* a meta-community). HCDT entropy and Hill numbers have been applied to phylogenetic and functional diversity, and, though always perfectible, solutions have been proposed to most issues: the correspondence between HCDT entropy and Hill numbers [17, 18], the decomposition of diversity according to Jost's [16, 19] or Routledge's definition [18, 20] of α diversity and robust estimators [21–24].

Recently, a major step forward was made by Leinster and Cobbold [25] who introduced a general measure of diversity denoted ${}^q\mathbf{D}^{\mathbf{Z}}$, the *similarity-based diversity*. Encompassing all previously cited measures, it allows a direct measure of diversity as the inverse of the average ordinariness of species. The latter is defined as its average similarity with other species: the topology of species, *i.e.* the place of this species in a multivariate space, does rely on species-to-species similarity. Parameterizing the definition of the average [26] allows choosing the importance of rare species. The great improvement of this measure is to build on a species-to-species distance matrix directly. This is a very desirable property that allows taking into account not only evolutionary or functional

distances but also any other kind of paired relationships [25]. Let's focus on functional diversity. While evolutionary relationships are naturally represented by a phylogenetic tree, functional diversity is more generally calculated from a matrix of multivariate distances, which is often highly distorted later when transformed into a functional dendrogram [19]. In addition to being problematic for the calculation of the functional diversity, this over-simplification of the raw trait matrix is extremely frustrating for field ecologists knowing the many difficulties encountered to measure these batteries of functional traits, especially in hyper-diverse ecosystems [27]. Knowing that single trait values or single trait variation axis are often inadequate predictors of the species fundamental niche (McGill et al. 2006), losing the multivariate information when estimating diversity prevents taking into account unique combinations of traits that influence ecosystems that are responsible for the ecosystem effect of a given species [28]. Up today, this *similarity-based diversity* ${}^qD^{\mathbf{Z}}$ still lacks both (i) a robust estimator and (ii) a decomposition framework. First, it has long been recognized that the observed diversity depends on the sample size [29], so that estimation-bias corrected estimators are required [30, 31]. Second, Leinster and Cobbold have said very little about ${}^qD^{\mathbf{Z}}$ decomposition into α, β and γ components. If the path is relatively well marked for diversity indices based on dendrograms, it remains to extend this analysis to the similarity-sensitive ones.

In this paper, we first recall the definitions and duality between similarity-based diversity and Ricotta and Szeidl's [32] entropy ${}^qH^{\mathbf{Z}}$. We propose two estimation-bias corrected estimators: the first one is an implementation of the Horvitz-Thompson [33] estimator, the other is built on the estimation of the ordinariness of unobserved species. Then we derive the decomposition of both ${}^qD^{\mathbf{Z}}$ and ${}^qH^{\mathbf{Z}}$ and provide a definition of β diversity and entropy.

2. Methods

2.1 Notations

Consider a random sample taken from a meta-community made of several local communities. Abundances of species in each local community is denoted $n_{s,i}$ ($s = 1, 2, \ldots, S$ is the index of species, i the index of communities). n_s is the number of individuals of species s in the meta-community, n_i the number of individuals sampled in local community i and n the total number. The same notations are used for probabilities of occurrence $p_{s,i}$ which are unknown but estimated with $\hat{p}_{s,i} = n_{s,i}/n_i$. Community weights are w_i: they may be equal to n_i/n but any positive values summing to 1 are allowed. We assume that $p_s = \sum_i w_i p_{s,i}$ for all species. This may be understood as the definition of our meta-community: the assemblage of communities whose species probabilities are the weighted average of those of communities. Diversity of the meta-community is γ diversity. Diversity of local communities is α diversity.

Species similarity is introduced as a square matrix $\mathbf{Z}$ of dimension $S \times S$ whose elements $z_{s,t}$ are the similarity between species s and t. Similarity is between 0 and 1 by definition, and $z_{s,s} = 1$: any species is completely similar to itself. A detailed presentation of the possible matrices can be found in [25] and is not repeated here. A matrix of particular interest is $\mathbf{Z} = \mathbf{I}_S$, the identity matrix where each species is completely different from the others, used to measure neutral diversity. The definition of ${}^qD^{\mathbf{Z}}$ can be extended to relatedness matrices, which only require that all terms $z_{s,t}$ are positive, and the diagonal terms strictly positive (see appendix A5 of [25]). Our results remain valid for relatedness matrices except for our new estimator which explicitly supposes that a species similarity with itself is 1.

2.2 Definition of ${}^qD^Z$ and ${}^qH^Z$

A species ordinariness is defined as its average similarity in the community:

$$(\mathbf{Zp})_s = \sum_t p_t z_{s,t} \tag{1}$$

${}^qD^{\mathbf{Z}}$ is the inverse of the generalized mean of order $q-1$ [26] of the community's species ordinariness:

$${}^qD^{\mathbf{Z}} = \left(\sum_s p_s (\mathbf{Zp})_s^{q-1} \right)^{\frac{1}{1-q}}; q \neq 1 \tag{2}$$

$${}^1D^{\mathbf{Z}} = \frac{1}{\prod_s (\mathbf{Zp})_s^{p_s}} \tag{3}$$

In other words, $1/{}^qD^{\mathbf{Z}}$ is the average ordinariness of species of the community. The generalized mean allows giving more or less importance to rare species.

${}^qD^{\mathbf{Z}}$ is the deformed exponential of order q of the entropy of Ricotta and Szeidl [32]:

$${}^qH^{\mathbf{Z}} = \frac{1 - \sum_s p_s (\mathbf{Zp})_s^{q-1}}{q-1} = \sum_s p_s \ln_q \frac{1}{(\mathbf{Zp})_s}; q \neq 1 \tag{4}$$

$${}^1H^{\mathbf{Z}} = -\sum_s p_s \ln (\mathbf{Zp})_s \tag{5}$$

Diversity is a monotonic, increasing function of entropy, which is more easily written with the formalism of deformed logarithms [34]. For $q \neq 1$:

$$\ln_q x = \frac{x^{1-q} - 1}{1-q} \tag{6}$$

$$e_q^x = [1 + (1-q)x]^{\frac{1}{1-q}} \tag{7}$$

The relation between entropy and diversity is the same as that of HCDT entropy and Hill numbers:

$$\ln_q {}^qD^{\mathbf{Z}} = {}^qH^{\mathbf{Z}} \tag{8}$$

$${}^qD^{\mathbf{Z}} = e_q^{{}^qH^{\mathbf{Z}}} \tag{9}$$

2.3 Estimation bias correction

As many other measures of diversity, ${}^qD^{\mathbf{Z}}$ suffers estimation bias [35]. The plug-in estimator, built by simply plugging $\hat{p}_s$ into the formula of ${}^qD^{\mathbf{Z}}$, is biased for two reasons: unobserved species [22] and the nonlinearity of diversity with respect to probabilities [36].

Bias-corrected estimators exist for HCDT entropy [20], including the Horvitz and Thompson [33] estimator first adapted by Chao and Shen [22] for Shannon entropy. We use it here to estimate ${}^qH^{\mathbf{Z}}$. We then propose a new estimator of ${}^qD^{\mathbf{Z}}$.

2.3.1 Chao-Shen estimator

${}^qH^{\mathbf{Z}}$ is the sum over species of the measure $h_s = p_s \ln_q \frac{1}{(\mathbf{Zp})_s}$. Unobserved species are responsible for unobserved terms of the sum. An unbiased estimator of such a sum has been derived by Horvitz and Thompson [33]: each observed term is divided by the probability for the species to be sampled: the lower it is, the higher h_s is weighted in the sum.

Chao and Shen [22] proposed to combine it with the estimator of the sample coverage. The total probability of occurrence of unobserved species is by definition 1 minus the sample coverage [37], which can be estimated from the data following Zhang and Huang [38]:

$$\hat{C} = 1 - \sum_{r=1}^{n} (-1)^{r+1} \binom{n}{r}^{-1} f_r \tag{10}$$

n is the sample size and f_r the number of species observed r times in the sample. This is an improvement of the well-known Turing's formula $\hat{C} = 1 - f_1/n$. The observed species probabilities can be better estimated by $\left\{\hat{C}\hat{p}_1, \ldots, \hat{C}\hat{p}_T\right\}$ where T is the number of observed species. $1-\hat{C}$ is the estimated total probability of unobserved species.

We also have to estimate $(\mathbf{Zp})_s$. The observed $\mathbf{Z}$ matrix lacks the $S-T$ columns of unobserved species, and the vector of probabilities lacks the corresponding probabilities $\{\hat{p}_{T+1}, \ldots, \hat{p}_S\}$. Nothing is known about the similarity of unobserved species with observed ones. A reasonable assumption is that their average similarity with any other species is identical to the average similarity between observed species, $\bar{z} = \left(\sum_{s,t\neq s} z_{s,t}\right)/[T\,(T-1)]$. Since the sum of missing probabilities of species is known to be $1-\hat{C}$, an estimator of $(\mathbf{Zp})_s$ is:

$$\left(\widetilde{\mathbf{Zp}}\right)_s = \sum_t \hat{C}\hat{p}_t z_{s,t} + \left(1-\hat{C}\right)\bar{z} \tag{11}$$

The first term is the sum over observed species of their similarity with species s and the second term is the expected sum over unobserved species. Plugging the estimator of p_s and that of $(\mathbf{Zp})_s$ into the Horvitz-Thompson estimator, an estimation-bias corrected estimator of ${}^qH^{\mathbf{Z}}$ is:

$${}^q\tilde{H}^{\mathbf{Z}} = \sum_s \frac{\hat{C}\hat{p}_s \ln_q \frac{1}{\left(\widetilde{\mathbf{Zp}}\right)_s}}{1-\left(1-\hat{C}\hat{p}_s\right)^n} \tag{12}$$

The estimator of diversity is ${}^q\tilde{D}^{\mathbf{Z}} = e_q^{{}^q\tilde{H}^{\mathbf{Z}}}$.

2.3.2 Alternative estimator

Instead of correcting the estimator for unobserved species, we now want to explicitly estimate their contribution to the value of diversity.

The observed $\mathbf{Z}$ matrix lacks $S-T$ lines and columns. The actual $\mathbf{Z}$ matrix is (only bold elements are known):

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1} & \cdots & \mathbf{z_{1,T}} & z_{1,T+1}=\bar{z} & \cdots & z_{1,S}=\bar{z} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \mathbf{z_{T,1}} & \cdots & \mathbf{1} & \bar{z} & & \bar{z} \\ z_{T+1,1}=\bar{z} & \cdots & \bar{z} & 1 & & \bar{z} \\ \vdots & & & & \ddots & \vdots \\ z_{S,1}=\bar{z} & \cdots & \bar{z} & \bar{z} & \cdots & 1 \end{pmatrix} \tag{13}$$

If $q \neq 1$, diversity can split into two terms:

$${}^qD^{\mathbf{Z}} = \left[\sum_{s\leq T} p_s (\mathbf{Zp})_s^{q-1} + \sum_{t>T} p_t (\mathbf{Zp})_t^{q-1}\right]^{\frac{1}{1-q}} = (K+U)^{\frac{1}{1-q}} \tag{14}$$

The first term K is easy to estimate, as in the Chao-Shen estimator:

$$\hat{K} = \sum_{s\leq T} \hat{C}\hat{p}_s \left[\left(\sum_{t\leq T} \hat{C}\hat{p}_t z_{s,t}\right) + \left(1-\hat{C}\right)\bar{z}\right]^{q-1} \tag{15}$$

The second one is more problematic:

$$U = \sum_{t>T} p_t [\bar{z}(1-p_t) + p_t]^{q-1} \tag{16}$$

The number of terms is unknown, as is each p_t, but U can be estimated following Zhang and Grabchak [24]. We define:

$$V = \sum_s p_s [\bar{z}(1-p_s) + p_s]^{q-1} \tag{17}$$

V is a linear diversity index, using the terminology of Zhang and Grabchak. It can be estimated (derivation in Appendix S1) by:

$$\hat{V} = 1 + \sum_{s\leq T} \frac{n_s}{n} \sum_{v=1}^{n-n_s} (1-\bar{z})^v \left[\prod_{i=1}^{v} \frac{i-q}{i}\right] \left[\prod_{j=1}^{v} \left(1 - \frac{n_s - 1}{n-j}\right)\right]$$

(18)

U is the sum of the terms concerning unobserved species, so observed terms must be subtracted from $\hat{V}$. The improved estimator of ${}^{q}D^{\mathbf{Z}}$ is thus:

$$ {}^{q}\tilde{D}^{\mathbf{Z}} = \left\{ \hat{K} + \hat{V} - \sum_{s \leq T} \hat{C}\hat{p}_s \left[\bar{z}\left(1 - \hat{C}\hat{p}_s\right) + \hat{C}\hat{p}_s \right]^{q-1} \right\}^{\frac{1}{1-q}} \quad (19) $$

The estimator of ${}^{q}H^{\mathbf{Z}}$ is:

$$ {}^{q}\tilde{H}^{\mathbf{Z}} = \frac{\hat{K} + \hat{V} - \sum_{s \leq T} \hat{C}\hat{p}_s \left[\bar{z}\left(1 - \hat{C}\hat{p}_s\right) + \hat{C}\hat{p}_s \right]^{q-1} - 1}{1-q} \quad (20) $$

If $q = 1$, estimating entropy is easier. It can be split into two terms:

$$ {}^{1}H^{\mathbf{Z}} = -\sum_{s \leq T} p_s \ln\left(\mathbf{Z}\mathbf{p}\right)_s - \sum_{t > T} p_t \ln\left(\mathbf{Z}\mathbf{p}\right)_t = L + X \quad (21) $$

As above, the first term is estimated from the data:

$$ \hat{L} = -\sum_{s \leq T} \hat{C}\hat{p}_s \ln\left[\left(\sum_{t \leq T} \hat{C}\hat{p}_j z_{s,t} \right) + \left(1 - \hat{C}\right)\bar{z} \right] \quad (22) $$

The second term addresses unobserved species:

$$ X = -\sum_{t gt T} p_t \ln\left[\bar{z}\left(1 - p_t\right) + p_t\right] \quad (23) $$

We define the linear diversity index:

$$ W = -\sum_{s} p_s \ln\left[\bar{z}\left(1 - p_s\right) + p_s\right] \quad (24) $$

W can be estimated by (Appendix S1):

$$ \hat{W} = \sum_{s \leq T} \frac{n_s}{n} \sum_{v=1}^{n-n_s} \frac{\left(1-\bar{z}\right)^v}{v} \left[\prod_{j=1}^{v} \left(1 - \frac{n_s - 1}{n - j} \right) \right] \quad (25) $$

Finally:

$$ {}^{1}\tilde{H}^{\mathbf{Z}} = \hat{L} + \hat{W} + \sum_{s \leq T} \hat{C}\hat{p}_s \ln\left[\bar{z}\left(1 - \hat{C}\hat{p}_s\right) + \hat{C}\hat{p}_s \right] \quad (26) $$

$$ {}^{1}\tilde{D}^{\mathbf{Z}} = e^{{}^{1}\tilde{H}^{\mathbf{Z}}} \quad (27) $$

2.4 Decomposition of entropy

We decompose entropy and diversity following Marcon *et al.* [18]. Entropy and diversity can be measured in each community. When communities are pooled, a meta-community (whose species probabilities are the weighted average of those of communities) is defined. Entropy and diversity of the meta-community are called γ. The α entropy of the meta-community is the weighted average entropy of communities, *i.e.* we follow Routledge's [39] definition of α entropy. Its deformed exponential is α diversity (which is thus not the average diversity of communities). Entropy is decomposed additively, according to Patil and Taillie's [8] concept of diversity of a mixture. Diversity is decomposed multiplicatively, and both decompositions are equivalent. Our purpose here is to characterize β entropy, beyond defining it only as the difference between γ and α entropy.

The α entropy of community i is:

$$ {}^{q}_{i}H^{\mathbf{Z}}_{\alpha} = \frac{1 - \sum_s p_{si} \left(\mathbf{Z}\mathbf{p}\right)_{si}^{q-1}}{q-1} \quad (28) $$

The α entropy of the meta-community is the weighted average entropy of communities:

$$ {}^{q}H^{\mathbf{Z}}_{\alpha} = \sum_i w_i \, {}^{q}_{i}H^{\mathbf{Z}}_{\alpha} = \frac{1 - \sum_i w_i \sum_s p_{si} \left(\mathbf{Z}\mathbf{p}\right)_{si}^{q-1}}{q-1} \quad (29) $$

Algebra detailed in Appendix S2 gives the expression of β entropy:

$$ {}^{q}H^{\mathbf{Z}}_{\beta} = \sum_i w_i \sum_s p_{si} \left(\ln_q \frac{1}{\left(\mathbf{Z}\mathbf{p}\right)_s} - \ln_q \frac{1}{\left(\mathbf{Z}\mathbf{p}\right)_{si}} \right) \quad (30) $$

β entropy is the generalized Jensen-Shannon divergence [20] between the distribution of $\mathbf{Z}\mathbf{p}$ in each community and in the meta-community.

The α diversity of the meta-community is the deformed exponential of its α entropy:

$$ {}^{q}D^{\mathbf{Z}}_{\alpha} = e_q^{{}^{q}H^{\mathbf{Z}}_{\alpha}} = \left[\sum_i w_i \sum_s p_{si} \left(\mathbf{Z}\mathbf{p}\right)_{si}^{q-1} \right]^{\frac{1}{1-q}} \quad (31) $$

β diversity can be obtained by taking the deformed exponential of the decomposition of entropy:

$$ {}^{q}D^{\mathbf{Z}}_{\beta} = e_q^{\frac{{}^{q}H^{\mathbf{Z}}_{\beta}}{1+(1-q){}^{q}H^{\mathbf{Z}}_{\alpha}}} \quad (32) $$

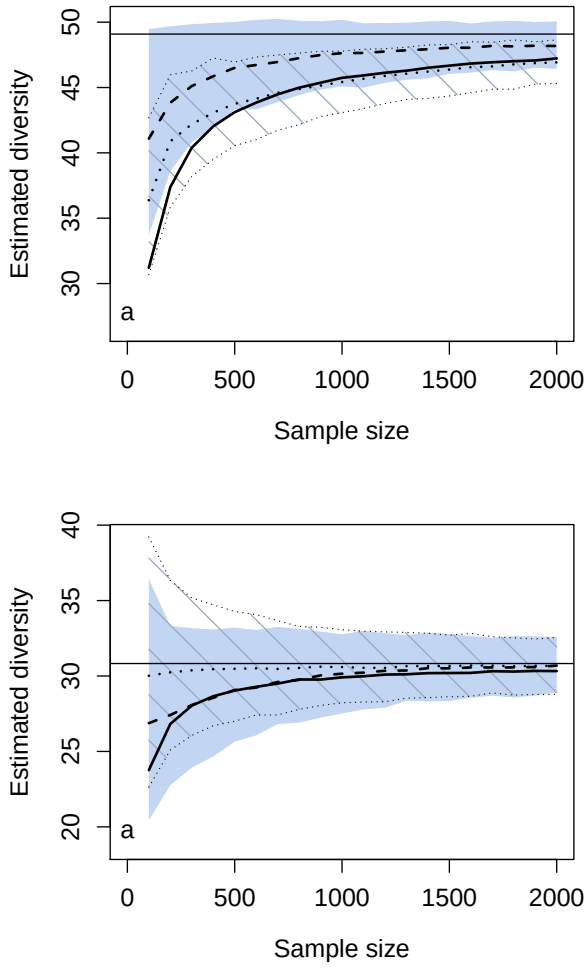

Figure 1. Estimation of BCI diversity ${}^{0.5}D^{\mathbf{Z}}$ (a) and ${}^{1.5}D^{\mathbf{Z}}$ (b) depending on sample size.
The plug-in estimator (plain line) is the least effective. The Chao-Shen estimator (dashed line, grey-shaded envelope) performs better than our new estimator (dotted line, dashed envelope) for small values of q but our new estimator is very effective for high values of q. Envelopes are the 2.5% and 97.5% quantiles of simulated values. The horizontal line is the diversity of the whole data set.

2.5 Decomposition of diversity

It is also interesting to directly decompose diversity to characterize β diversity. It is usually only defined as the ratio between γ and α diversity [40] or as a transformation of β entropy [18, 20, 41].

The inverse of ${}^{q}D_{\gamma}^{\mathbf{Z}}$ is the generalized mean of order $q-1$ of $(\mathbf{Zp})_s$:

$$\frac{1}{{}^{q}D_{\gamma}^{\mathbf{Z}}} = \left[\sum_i p_s (\mathbf{Zp})_s^{q-1}\right]^{\frac{1}{q-1}} \tag{33}$$

From ${}^{q}D_{\gamma} = {}^{q}D_{\alpha}\,{}^{q}D_{\beta}$ and Routledge weighting of alpha diversity, simple algebra (Appendix S3) yields:

$${}^{q}D_{\beta}^{\mathbf{Z}} = \left[\sum_i w_i \left(\frac{1/{}_i^{q}D_{\alpha}^{\mathbf{Z}}}{1/{}^{q}D_{\gamma}^{\mathbf{Z}}}\right)^{q-1}\right]^{\frac{1}{q-1}} \tag{34}$$

As $1/{}^{q}D_{\gamma}^{\mathbf{Z}}$ is the average ordinariness of species, ${}^{q}D_{\beta}^{\mathbf{Z}}$ is the average normalized ordinariness of communities where the normalized ordinariness of community i is defined as $\frac{1/{}_i^{q}D_{\alpha}^{\mathbf{Z}}}{1/{}^{q}D_{\gamma}^{\mathbf{Z}}}$, that is to say its average species ordinariness divided by that of the meta-community.

We may consider each community as an assemblage of monospecific populations. Then, each population's diversity is ${}_s^{q}D_{\alpha}^{\mathbf{Z}} = 1$, and weights are p_s. Introducing these values into eq. (34) yields ${}^{q}D_{\gamma}^{\mathbf{Z}} = {}^{q}D_{\beta}^{\mathbf{Z}}$: as shown by Rao and Nayak [42] for entropy, the diversity of an assemblage of species is the β diversity between its monospecific assemblages.

3. Results

3.1 Test of bias correction

We used the Barro-Colorado Island (BCI) 50-ha plot forest inventory [43–45]. Year 2005 census contains 20852 individuals from 229 species, among which 24 have been observed only once. The sample coverage is close to 99.9%, allowing to consider that the inventory is almost exhaustive and to use it as a reference to test the efficiency of estimators applied to subsamples. We set species similarity equal to 2/3 inside a genus, 1/3 inside a family, and 0 outside family. The average similarity between pairs of distinct species is $\bar{z} \approx 0.01$. The similarity matrix is somehow rough but no better data was available for such a number of species.

We simulated inventories between 100 and 2000 individuals in a multinomial distribution respecting same species probabilities. Simulations were repeated 1000 times, estimated entropy averaged and finally transformed into diversity for plotting.

Simulated estimations are plotted in Figure 1. 95% confidence envelopes are shown: simulated samples are realizations of the assumed multinomial distribution of the community; stochasticity is not due to the estimators but to sampling.

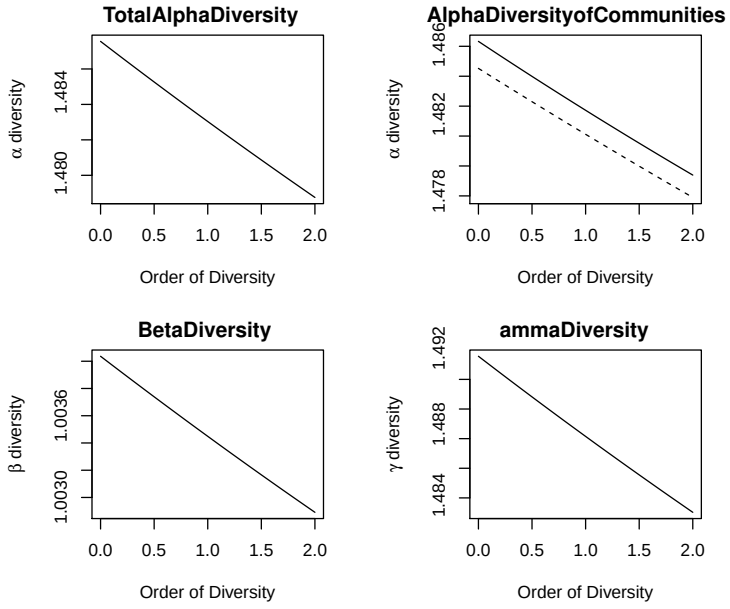

Figure 2. Functional diversity profile of the Paracou forest communities.
Top left: α diversity of the meta-community; top right: α diversity of each community (P006: plain line, P018: dotted line); bottom left: β diversity; bottom right: γ diversity. Diversity is the number of effective species (effective communities for diversity) against the order q. It is estimated with bias correction.

The best estimator depends on the value of q. Their correction is almost identical at $q = 1.2$ (not shown). At $q = 1.5$, diversity can be estimated very accurately by the new estimator with a sample whose size is less than the number of species of the community. Following Marcon *et al.* [20]; we choose a pragmatic estimation-bias correction using the maximum value of the two estimators.

3.2 Partitioning functional diversity

We used the same tropical forest dataset as Marcon and Hérault [18], made of two 1-ha fully inventoried plots in the Paracou field station in French Guiana. 1124 individual trees (diameter at breast height over 10 cm) have been sampled among 229 species. Four key functional traits were addressed: seed mass and tree maximum height [46], and specific leaf area and wood specific gravity [27]. A dissimilarity matrix was first built using the Gower metric by the *daisy* function of the cluster package [47] for R [48]. The similarity matrix $\mathbf{Z}$ was defined as 1 minus the normalized dissimilarity matrix (*i.e.* all values were divided by their maximum), so that its values are between 0 (the less similar pair of species) and 1 (the diagonal of the matrix).

We calculated diversity of order 1. Neutral γ diversity is 134 effective species. Similarity-based functional γ diversity is 1.49 effective species. It can be interpreted as: the average species ordinariness is $1/1.49 = 0.67$. It is also a Hill number: 1.49 completely different species with equal frequencies would have the same diversity as the observed meta-community made of 229 species. This is a very small value; due to the low functional distances between species in the similarity matrix: $\bar{z} \approx 0.65$.

Neutral β diversity is 1.46 effective communities: the species distributions are very different between the communities. Yet, functional β diversity is 1.003: almost no functional β diversity is detected between the communities. Figure 2 shows the functional diversity profile of the meta-community: its values are almost constant whatever q.

4. Discussion and Conclusion

In this paper, we derived the decomposition of similarity-based diversity ${}^{q}D^{\mathbf{Z}}$ and proposed two estimators to reduce estimation bias from the level of the plug-in estimator. The Chao-Shen estimator is the best for small orders of diversity but our alternative estimator outperforms it for higher orders.

Similarity-based diversity is preferred to estimate functional diversity from a distance matrix because it does not require building a dendrogram and so it preserves the topology of species in the space of functional traits. Even if we focused mainly on the functional diversity, the method works equally well with any similarity measures. As highlighted by Leinster and Cobbold, similarity can be measured in any meaningful way: a genetic similarity will lead to a decomposition of genetic diversity; a molecular similarity will lead to a decomposition of molecular diversity, and so on. Results are effective number of basic entities and entity assemblages which have the same desirable properties as classical diver-

sity measures.

We estimated functional β diversity between two tropical forest plots previously investigated by Marcon and Hérault [18] with different methods. Our results confirmed and amplified the previous conclusions: functional β diversity between these two quite similar forest communities is negligible (very close to 1). The main explanation is probably the high functional redundancy between these two forest plots. The latter are spaced by only a few hundred meters and are located on similar topography and soils. It is always strange to observe how tropical forests can locally be extremely rich in species but exhibit a very low turnover in space [49, 50]. This finding, well-documented for neutral diversity, seems here exacerbated from the perspective of the functional diversity. This supports the view that, in tropical forests, niche differentiation, if any, occur at very fine spatial scale through vertical distribution shaped by light access and micro-habitat complementarity [51]. Above this micro-habitat scale, functional assemblages are very similar to each other and governed by a few leading and continuous functional axes [27]. This is supported by the flat shape of diversity profiles: giving more importance to rare species (lowering q) has no effect because rare species are not less ordinary. In other words, according to our limited data and number of traits, functional diversity is brought by frequent species and quickly saturated: adding more species or more communities does not increase it.

The entropart package [52] for R, starting from version 1.2, allows diversity estimation presented in the paper.

Acknowledgments

This work has benefited from an “Investissement d’Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01). Funding came from the project Climfor (Fondation pour la Recherche sur la Biodiversité).

References

[1] Gallopín GC (2006) Linkages between vulnerability, resilience, and adaptive capacity. Global Environmental Change 16: 293–303.

[2] Violle C, Navas ML, Vile D, Kazakou E, Fortunel C, et al. (2007) Let the concept of trait be functional! Oikos 116: 882–892.

[3] Weiner J (1995) On the practice of ecology. Journal of Ecology 83: 153–158.

[4] Shannon CE (1948) A Mathematical Theory of Communication. The Bell System Technical Journal 27: 379–423, 623–656.

[5] MacArthur R (1955) Fluctuations of Animal Populations and a Measure of Community Stability. Ecology 36: 533–536.

[6] Hurlbert SH (1971) The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology 52: 577–586.

[7] Rényi A (1961) On Measures of Entropy and Information. In: Neyman J, editor, 4th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: University of California Press, volume 1, pp. 547–561.

[8] Patil GP, Taillie C (1982) Diversity as a concept and its measurement. Journal of the American Statistical Association 77: 548–561.

[9] Dalton H (1920) The measurement of the inequality of incomes. The Economic Journal 30: 348–361.

[10] Havrda J, Charvát F (1967) Quantification method of classification processes. Concept of structural a-entropy. Kybernetika 3: 30–35.

[11] Daróczy Z (1970) Generalized information functions. Information and Control 16: 36–51.

[12] Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. Journal of Statistical Physics 52: 479–487.

[13] MacArthur RH (1965) Patterns of species diversity. Biological Reviews 40: 510–533.

[14] Adelman MA (1969) Comment on the "H" Concentration Measure as a Numbers-Equivalent. The Review of Economics and Statistics 51: 99–101.

[15] Hill MO (1973) Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology 54: 427–432.

[16] Jost L (2007) Partitioning diversity into independent alpha and beta components. Ecology 88: 2427–2439.

[17] Jost L (2006) Entropy and diversity. Oikos 113: 363–375.

[18] Marcon E, Hérault B (2014) Decomposing Phylodiversity. HAL hal-009461: 1–22.

[19] Chiu CH, Jost L, Chao A (2014) Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. Ecological Monographs 84: 21–44.

[20] Marcon E, Scotti I, Hérault B, Rossi V, Lang G (2014) Generalization of the partitioning of Shannon diversity. Plos One 9: e90289.

[21] Grassberger P (1988) Finite sample corrections to entropy and dimension estimates. Physics Letters A 128: 369–373.

[22] Chao A, Shen TJ (2003) Nonparametric estimation of Shannon’s index of diversity when there are unseen species in sample. Environmental and Ecological Statistics 10: 429–443.

[23] Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, et al. (2013) Robust estimation of microbial diversity in theory and in practice. The ISME journal 7: 1092–101.

[24] Zhang Z, Grabchak M (2014) Entropic Representation and Estimation of Diversity Indices. arXiv 1403.3031: 1–12.

[25] Leinster T, Cobbold C (2012) Measuring diversity: the importance of species similarity. Ecology 93: 477–489.

[26] Hardy GH, Littlewood JE, Pólya G (1952) Inequalities. Cambridge University Press.

[27] Baraloto C, Paine CETP, Patiño S, Bonal D, Hérault B, et al. (2010) Functional trait variation and sampling strategies in species rich plant communities. Functional Ecology 24: 208–216.

[28] Eviner VT, Chapin FSI (2003) Functional Matrix: A Conceptual Framework for Predicting Multiple Plant Effects on Ecosystem Processes. Annual Review of Ecology, Evolution, and Systematics 34: 455–485.

[29] Wolda H (1981) Similarity indices, sample size and diversity. Oecologia 50: 296–302.

[30] Beck J, Holloway JD, Schwanghart W (2013) Undersampling and the measurement of beta diversity. Methods in Ecology and Evolution 4: 370–382.

[31] Butturi-Gomes D, Junior MP, Giacomini HC, Junior PDM (2014) Computer intensive methods for controlling bias in a generalized species diversity index. Ecological Indicators 37, Part A: 90–98.

[32] Ricotta C (2006) Strong requirements for weak diversities. Diversity and Distributions 12: 218–219.

[33] Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47: 663–685.

[34] Tsallis C (1994) What are the numbers that experiments provide? Química Nova 17: 468–471.

[35] Dauby G, Hardy OJ (2012) Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species. Ecography 35: 661–672.

[36] Bonachela JA, Hinrichsen H, Muñoz MA (2008) Entropy estimates of small data sets. Journal of Physics A: Mathematical and Theoretical 41: 1–9.

[37] Good IJ (1953) On the Population Frequency of Species and the Estimation of Population Parameters. Biometrika 40: 237–264.

[38] Zhang Z, Huang H (2007) Turing's formula revisited. Journal of Quantitative Linguistics 14: 222–241.

[39] Routledge RD (1979) Diversity indices: Which ones are admissible? Journal of Theoretical Biology 76: 503–515.

[40] Tuomisto H (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography 33: 2–22.

[41] Marcon E, Puech F, Traissac S (2012) Characterizing the relative spatial structure of point patterns. International Journal of Ecology 2012: 11.

[42] Rao C, Nayak T (1985) Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. Information Theory, IEEE Transactions on 31: 589–593.

[43] Condit R (1998) Tropical Forest Census Plots. Berlin, Germany, and Georgetown, Texas: Springer-Verlag and R. G. Landes Company, 1–224 pp.

[44] Hubbell SP (1999) Light-Gap Disturbances, Recruitment Limitation, and Tree Diversity in a Neotropical Forest. Science 283: 554–557.

[45] Hubbell SP, Condit R, Foster RB (2005). Barro Colorado Forest Census Plot Data.

[46] Hérault B, Bachelot B, Poorter L, Rossi V, Bongers F, et al. (2011) Functional traits shape ontogenetic growth trajectories of rain forest tree species. Journal of Ecology 99: 1431–1440.

[47] Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2014). cluster: Cluster Analysis Basics and Extensions.

[48] R Development Core Team (2014). R: A Language and Environment for Statistical Computing. URL http://www.r-_project.org.

[49] Condit R, Pitman N, Leigh EG, Chave J, Terborgh J, et al. (2002) Beta-diversity in tropical forest trees. Science 295: 666–669.

[50] Novotny V, Miller SE, Hulcr J, Drew RaI, Basset Y, et al. (2007) Low beta diversity of herbivorous insects in tropical forests. Nature 448: 692–5.

[51] Kraft NJB, Valencia R, Ackerly DD (2008) Assembly in an Amazonian Forest Tree Community Functional Traits and Niche-Based. Science 322: 580–582.

[52] Marcon E, Hérault B (2014). entropart, an R package to partition diversity. URL http://cran.r-_project.org/package=entropart.